中国大学生的网络使用:基于大规模日志分析的模式识别新方法

■ 严承希 王军 王珂
北京大学信息管理系 北京 100871

摘要:[**目的/意义**]深入挖掘和准确理解中国大学生日常网络行为模式,不仅对促进用户行为和检索领域的发展具有巨大的理论意义,而且在提升面向大学生用户的企业个性化服务与信息推荐能力方面也具有潜在的社会价值和实践意义。[**方法/过程**]提出一种基于大规模日志分析的大学生用户行为模式识别新方法,该方法包括一种基于深度学习和文本分析技术的半监督学习算法"MaxMatching"以及混合两种特征熵(香农熵与真实熵)的聚类模型。[**结果/结论**]实证结果表明本方法不仅在算法和结果解释上具有一定的优势,而且能从网络使用能力、访问时序性和主题倾向性三方面归纳与呈现中国大学生网络行为全方位模式。该方法和结论有效地拓展了信息检索领域查询项的语义化理解方面的方法体系,也为企业提升面向大学生用户的个性化信息推荐服务提供一定的参考和可行性意见。

关键词:中国大学生 网络行为 模式识别 大规模日志分析
分类号:G250
DOI:10.13266/j.issn.0252-3116.2019.14.010

1 引言

信息技术的发展使网络成为人们生活层面的重要组成部分。根据 2018 年中国互联网络信息中心(China Internet Network Information Center, CNNIC)统计报告[1],20-29 岁年龄段网民通过 PC 电脑、手机终端等多种网络渠道使用互联网,占中国网民人数的 30%;大专、本科及以上学历占到 21.1%。不同于其他年龄段的人群,以"90 后"为主力的大学生群体在网络技术日益革新的时代中更加容易接受新文化、新思想以及新技术的传播与影响,具体表现在其日常网络生活中的行为模式与主题偏好中,如较强的网络搜索能力、对亚文化和游戏文化的追求与社交媒体的广泛使用。因此理解和搜寻有用的用户模式并识别出有意义的事件、潜在的风险和制定战略决策具有深远的社会意义[2]。大规模网络日志分析正是一种基于海量的用户网络行为记录,通过数据挖掘和机器学习算法对不同用户群体的宏观结构和微观特征进行逐层分析与揭示的高效方法。虽然如此,但日志分析方法的效用仍然没有得以充分发挥,H. R. Jamali 等认为尽管基于人口统计数据的用户分类研究取得一些进展,但它并不是一种发现用户不同主题之间的信息寻求行为差异的非常有效的技术[3]。对此,本研究提出一种基于大规模日志分析的大学生用户行为模式识别新方法,该方法包括一种基于面向大学生日志数据深度学习和文本分析技术的非监督学习算法"MaxMatching",以及混合两种特征熵(香农熵与真实熵)的聚类模型,力图解决两个研究问题:①如何对日志数据中的查询项进行语义分析并准确理解用户使用意图和主题偏好?②在综合考虑大学生上网行为的网络使用能力特征、时序特征和主题特征下,如何理解不同大学生群体的网络行为模式?

2 研究综述

已有的相关研究对大学生信息搜索与使用行为、主题偏好和社会心理变化等多方面进行了探索,如在线音乐使用[4]、网络使用行为与心理因素[5-6]、学习型搜寻行为[7]等。M. Madden 等研究用户在线行为时发现大学生喜欢下载和听音乐,也喜欢进行在线聊天与

作者简介:严承希(ORCID:0000-0003-1128-550X),博士研究生;王军(ORCID:0000-0003-2850-0624),教授,博士生导师,通讯作者,E-mail:junwang@pku.edu.cn;王珂(ORCID:0000-0002-9941-1664),硕士研究生。
收稿日期:2018-12-06 修回日期:2019-03-06 本文起止页码:83-93 本文责任编辑:易飞

社交,不过很少出于休闲娱乐的目的[8]。张鹏翼等调查了中国大学生日常使用移动设备进行个人信息管理的活动,发现越来越多的大学生使用手机进行个人信息存储,其中近一半的被调查者会使用和存取通话、照相、社交媒体、邮件、个人便签、时钟和工作或个人文档等信息[9]。吴丹等重点研究大学生使用手机搜索所引发的跟随行为,在非受控实验环境下中对 30 位大学生近 15 天的手机使用情况进行了记录,并且结合结构化日记和采访数据进行定性与定量相结合的综合分析,研究结果显示存在三类跟随行为,即持续性搜索、购物决策和信息分享,并且大部分跟随行为会在首次搜索会话后 1 个小时后发生,大部分参与者会根据不同的 App 采取不同策略跟随——只有当搜索反馈结果满足用户的需求时,用户才会进行后续购物和分享行为,否则用户将使用不同 App 或修改查询项进行再次搜索[10]。

在信息检索和使用过程中,基于日志分析的查询项语义理解问题一直以来是计算机和情报学等领域的研究重点。最为经典的用户意图分类方法来自于 A. Z. Broder 提出的 INT 分类法[11],包括"信息类意图"(informational)、"导航类意图"(navigational)和"事务类意图"(transactional)。O. Alonso 等通过基于查询项的众包标注发现信息类查询在当前查询项中可以占 90% 以上的比重[12]。C. Gonzalez-Caro 等将查询意图分为"信息类意图"(informational)、"非信息类意图"(not informational)与"歧义性意图"(ambiguous),并根据 N. J. Belkin 搜索任务情景理论提出一种查询意图多分面分类方法,即将用户查询意图分为包括类型、主题、任务、客观性、具现性、范围、权威敏感性、空间敏感性和时间敏感性在内的 9 种分面,其中任务分面与查询项的资源类型有关[13-14]。

近年来学者们注意到多维度查询项特征对于深度的用户查询意图分类任务的重要性,如返回结果记录[15]、查询项长度[16]、查询词的词性与位置特征[17-18]与鼠标浏览特征[19]等。R. V. Pujeri 等认为查询项产生歧义性的原因一般是搜索项长度过短以至于无法包含足够的知识背景[20],这一点与 H. Cao 等提出的观点"查询项需要情景化感知(context-aware query)"[21]是一致的。J. Teevan 等使用查询结果的质量清晰度、点击熵以及查询项自身的属性(查询字符长度、是否包含 URL 或者是否包含地理信息等)构建出一套多查询词特征的贝叶斯依存网络分类模型,实证结果表明其分类预测的准确性可以达到 80% 左右[22]。除此以外,基于 taxonomy 的半监督查询项分类[23-24]、LDA 主题建模[25-26]与深度神经网络模型[27-28]近些年来也取得较为突出的研究成果。S. Dou 等提出基于查询项映射桥接 taxonomy 的分类算法,该方法使用开放式分类目录(open directory project, ODP)作为中间 taxonomy,通过最大化查询项与分面词表之间的匹配得分函数获取二者之间的候选关系,然后基于支持向量机进行分类建模。实验证明与 ACM KDDCUP 2005 比赛的第一名算法相比,此方法分别在 F1 和 Precision 指标上可提高 3% 和 9% 左右[29]。T. KONISHI 等注意到 LDA 模型存在非稀疏性的强假设的局限,于是将查询项的主题对共现关系考虑到主题模型中,提出一种成对主题模型 PCTM[30]。该模型使用针对每个词语的成对主题共现概率来进行塌缩性吉布斯抽样(collapsed Gibbs sampling)以解决主题之间稀疏性关联问题。实验结果表明 PCTM 在查准率上比 LDA 等传统模型超出 3%。郭程等结合 Hownet 和 ATF * PDF 模型提出一种面向查询项的无指导的主题挖掘模型,该模型对部分词频较小但相对重要的主题词汇有很好的识别力[31]。B. Wu 等基于级联假设分别将点击页面和跳过页面作为正负反馈文档集合,结合页面的内容和位置嵌入向量并构建带有注意力机制的深度反馈记忆网络(feedback memory network),该模型在查询项提示任务以及不同长度和会话的查询项意图识别任务中都获得最优的评价效果[32]。另外还有一些其他的查询意图识别方法如查询子项意图分解[33-34]以及关键实体识别等技术[35-36]也在特定任务上具有一定优势。

针对上述研究进行分析与总结,我们认为:①有关大学生群体的网络行为研究主要采用局部的问卷调查或用户访谈方法,且侧重于用户在移动终端的使用行为与心理因素的分析。由于受到小规模数据量以及设备场景的限制,其结论与成果可能存在偏差(数据偏差和主观认知的偏差)。②在用户意图识别分析上,大量研究仍是以"Broder 用户意图分类"为基础的粗粒度的拓展,少有结合特定群体(大学生用户)的时间、主题和行为层次进行多维度在线行为模式分析,正如 Y. K. Seock 所说"网络对大学生生活的渗透已经改变了他们的行为、习惯和偏好等,而不仅仅是不同设备使用方式的问题"[37]。③在日志分析中如何准确理解查询项(query)语义和用户查询意图一直是信息检索和模式识别领域的研究难点,目前以 taxonomy、主题模型或者深度学习等为基础的查询扩展的方法主要是强监督学习算法,不仅依赖于大量的高质量标注训练样本,同

严承希，王军，王珂．中国大学生的网络使用：基于大规模日志分析的模式识别新方法[J]．图书情报工作，2019，63(14)：83－93．

时计算复杂度和实现都较为困难，其高代价成本可能使中小型企业难以在生产实践上进行使用与部署。因此设计简单有效的非监督（弱监督）学习模型并用以准确识别中国大学生网络行为模式是非常值得探索的问题。

3 研究方法

3.1 研究框架

本节主要说明基于大规模日志分析的大学生用户行为模式识别方法的框架与步骤，如图1所示。我们首先设计出一个面向大学生用户在线网络信息需求的原型导航网站（见图2），并将其投放和嵌入到覆盖若干省份的中国高校大学校园网关服务中，并搜集大学生用户的上网日志记录，包括用户登录时间、点击网站的url以及搜索使用的查询项等。考虑到大学生用户信息需求与兴趣偏好，我们构建了基于查询项与日志记录的主题分类表，对网站url进行人工语义化标引。针对搜索的查询项数据，本研究根据外部语料知识和机器学习理论设计出一种半监督匹配学习算法Max-Matching，实现查询词的主题映射与转换，与url转化的主题记录进行合并。通过引入“时间、行为和主题”组成的三元组的特征熵，本研究将用户在线行为进行特征表示与抽取，并基于聚类分析模型实现大学生用户群体行为模式的识别。

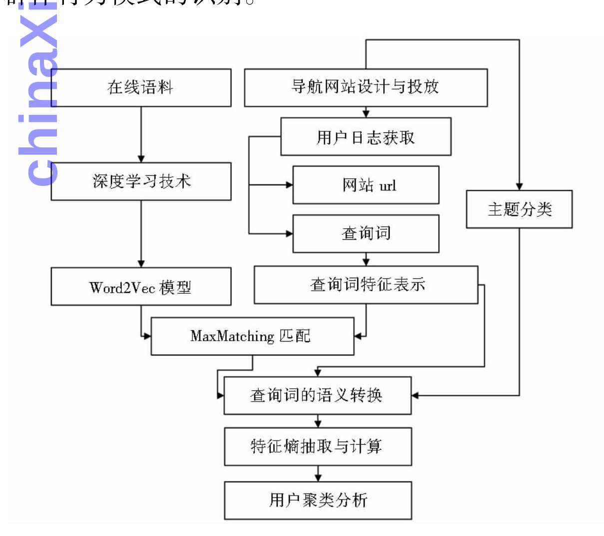


图1 基于大规模日志分析的大学生用户行为模式识别新方法的框架

3.2 导航网站设计与投放

为了获得大学生用户日常上网数据记录，本研究将Alexa 2016年网站排名靠前并覆盖大学生日常网络生活的9个方面（“吃”“玩”“乐”“挣”“聊”等）的网站

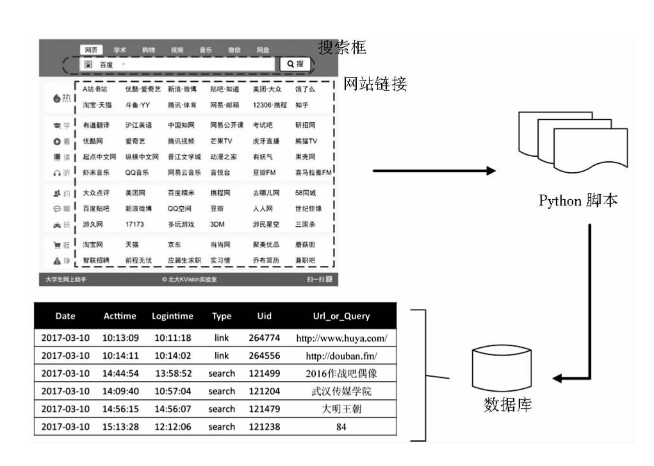


图2 导航网站的设计与数据搜集

进行筛选（共76个网站），从而构建和设计出一个用户友好的导航网站，进而联合网络代理商将该导航网站投放至全国不同省份，包括湖北、江西、广东、浙江、河北等20多个省份79个地市，覆盖全国近150所高校的大学生用户。为了确保该网站有较稳定的高使用率，网络代理商企业将该网站（包括数据使用的隐私协议等）嵌入到其所服务的各个高校校园网网关系统入口处，这些学校的大学生用户在登录校园网之后可以第一时间看到该网站并自由使用或者关闭此导航服务。同时根据我们对不同大学生用户使用记录和频次的统计，网络代理商将为高频使用此网站的用户提供上网套餐减免和优惠，以此来鼓励大学生们尽可能多地使用该导航网站。

本研究选择2017年3月10日到2018年3月10日整整一年的用户数据作为本研究的数据集，包括近3 500多个用户的40多万条日志记录。我们针对该数据集的使用情况进行统计后发现，网站运营指标独立用户访问量UV与网页浏览量PV分别达到每天平均36 897和73 727次，其转化率基本维持在3%左右的较高水平。从数据搜集的覆盖面和用户使用状况来看，我们认为该网站搜集的数据样本是可以一定程度代表中国大学生网络行为的。然后本研究通过对导航网站JS埋点和权威第三方平台百度统计采集大学生网站访问数据，并编写自动化R脚本进行定时的数据下载，存储于本地的数据库中，见图2。为了限定采样用户的范围，我们在python脚本中使用网关登录后的身份字符标识段（uid的前三位）进行过滤，以确定访问用户为在校大学生群体（包括在线本科生与研究生）。

3.3 用户日志预处理

用户日志的预处理阶段主要包含两部分的处理内

容:①对已有日志数据进行数据清洗,包括无效查询词的剔除、错误和遗漏的用户属性字段的过滤和错误条目的排除等,最终本研究共得到 3 550 名用户 347 387 条记录数据。数据字段包括 6 个,即用户账号(uid)、用户访问日期(date)、用户点击或检索行为的时间(acttime)、用户登录网站的时间(logintime)、用户行为类型(type)以及项目(item)。其中,用户行为类型包括搜索(search)或者链接点击(link),项目则包含查询词(query)或者网站 url,具体如图 2 所示。②根据已有的研究和前期调研情况,我们对大学生上网偏好和意图进行主题分类,采用人工标引对网站进行语义映射,并提供不同分类的概念词汇,以作为后续查询词语义匹配的种子词集合,该主题分类具体包括如表 1 所示:

表 1　用户日志的主题分类和网站映射

主题标签	主题类别	网站 url	种子词
Learning & Tool	学习工具类	有道翻译、网易公开课、中国知网、考试吧、沪江英语、研招网、我要自学网、智慧树	工具、邮箱、翻译、软件、大学、学习、考试、论文、编程、外语、课程、数据库
Job seeking	工作求职类	前程无忧、兼职吧、乔布简历、实习僧、应届生求职网、智联招聘、1010 兼职网	工作、求职、简历、兼职、实习、招聘
Art & Entertainment	文艺娱乐类	豆瓣 FM、网易云音乐、晋江文学城、起点中文网、虾米音乐、喜马拉雅 FM、音悦台、纵横中文网、QQ 音乐	文艺、文学、娱乐、阅读、音乐、时尚
Game & Animation	游戏动漫类	ACFun、哔哩哔哩、17173、3DM 游戏、动漫之家、多玩游戏、游民星空、三国杀、有妖气、游久网	游戏、动漫、二次元
Social communication	社交聊天类	网易邮箱、新浪微博、腾讯微博、人人、QQ 空间、百度贴吧、世纪佳缘	微博、博客、社交、交友、社区
Live video	视频直播类	斗鱼直播、直播吧、YY 直播、虎牙直播、腾讯视频、优酷网、爱奇艺、芒果 TV、熊猫 TV	视频、电影、TV、直播、电视剧
Comprehensive consumption	购物消费类	天猫、淘宝网、京东、当当、58 同城、美团网、大众点评、饿了么、聚美优品、去哪儿、糯米网、蘑菇街、12306	购物、消费、旅游、购书、点评、电商
knowledge & Information portal	资讯门户类	腾讯体育、网易、腾讯网、知乎、新浪网、百度知道、豆瓣、果壳	咨询、门户、知识、分享、问题、导航、分类

正如综述文献中所提到的,大部分查询项长度都比较短,且词项往往是非规范的自然语言,存在多歧义或新登录词等问题如“84”“跑男”以及“龙珠”等,难以被计算机系统所理解。本研究类似地采用查询扩展的策略,从人机交互的角度引入搜索引擎对查询项的返回高排序记录的元数据表达,这里主要选取前 top10 的记录作为查询项的背景语义知识。A. MALIK 等的研究表明对于绝大部分用户而言(特别是大学生用户),他们只对搜索引擎返回的前 10 个左右的网页记录满意[38]。为了对这些返回记录的元数据进行合理的语义化表达,本研究引入词向量的分布式嵌入表达方法,即基于深度神经网络 Word2Vec 模型[39]进行开放语料的预训练,然后设计出一种新的半监督启发匹配算法 MaxMatching,对查询词进行分类主题的识别与转换。在预训练阶段,我们从百度百科、搜狐新闻和搜狗语料爬取了近 13 000 000 个(130G)文本资源,通过 jieba 分词和 CBOW 模型进行词向量的训练(词窗口 window = 5,最小词频 min_count = 5,词特征维度 size of vector = 64),获取了涵盖 6 100 000 个词向量的超大词典。

MaxMatching 算法假设对于第 k 个 $query_k$ 的返回记录 t_i^k,其元数据的关键词 wr_p^i 与主题分类 s_j 的某个种子词 w_q^j 的平均语义相似度 W2V_similarity 可以作为该查询词的某类主题倾向性,并采用负指数权重进行加权(搜索引擎返回排在前面记录用户选择的可能性较大),最终计算出概率最大的主题类别即为该查询项所属的主题类别 $MM_RS_{j,k}$,具体流程和计算公式分别如图 3 和公式(1)所示:

$$MM_RS_{j,k} = \mathrm{argmax}_{j,k} \left\{ \sum_{i}^{T-1} e^{-i} \cdot \frac{1}{P \cdot Q} \sum_{p=0,q=0}^{P-1,Q-1} W2V_similarity(wr_p^i, w_q^j) \right\} \quad 式(1)$$

3.4　特征熵表示

本研究将构建出一个包含行为、时间和主题特征的三元组表达,记作 <‘behavior’,‘temporality’,‘topicality’>。已有研究已经证明了具有不同网络搜索能力的用户在点击和搜索使用习惯的差异,例如 D. Tabatabai 等的研究结果说明搜索能力较差的用户更多地倾向于无耐心的试错策略,这将直接导致他们在花费足够时间进行评估与计划之前更多地去选择和点击链接导航[40]。R. Mihalcea 则提出网络能力的概念(network competence)[41]来描述和刻画用户搜寻行为上的特点,如 ICT 工具使用偏好性。基于此,本研究将这种网络能力(或者行为上的使用偏好)记作 SC_{ratio},并

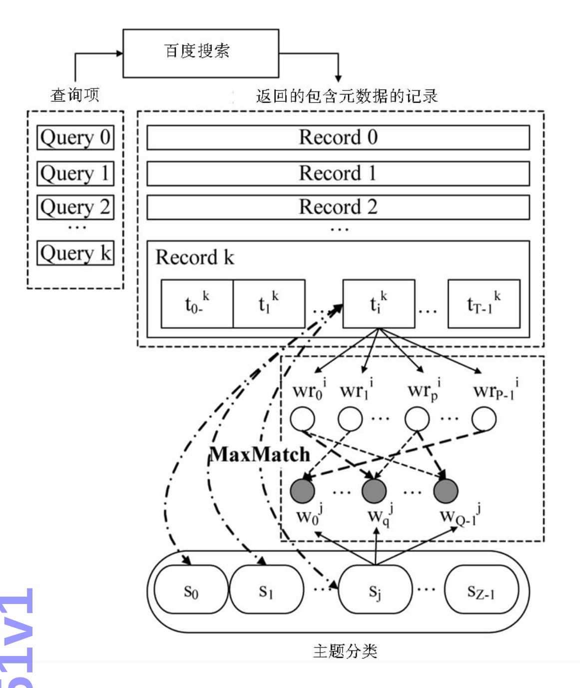


图 3　MaxMatching 算法流程

使用用户搜索频次在访问记录中的比率来予以衡量，如公式(2)所示：

$$SC_{ratio} = \frac{SeachingNum}{ClickingNum + SeachingNum} \quad 式(2)$$

信息熵(shannon entropy, SE)本质上用来刻画随机变量的不确定性，即我们对于信息的内容越不确定，则弄清楚它所需要的信息量越大，同理，用户在不同主题类中的选择可能性越相似，则其信息熵值越大，反映出用户不存在明确的主题倾向，这里我们用以衡量主题的专一性特征(topicality)。由于用户访问网络的时间是存在先后顺序的，仅仅使用信息熵来度量用户访问时间序列特征存在问题。A. Barabasi 等对此提出了真实熵(actual entropy, AE)这一概念[42]，有效地解决了序列先后的熵值预测问题。本研究使用 AE 对用户访问行为时间序列进行计算，来判断用户的访问有序或者规律程度。如果 AE 大，则说明用户访问行为的时间特征是无规律(无序)的。假设 $P(x_j)$ 是主题 x_j 的发生概率，φ_i 表示起始于时间序列位置 i 但没有在位置 1 到 i – 1 中出现过的最短的串长长度，Z 和 n 分别表示独立用户访问的主题类数和序列串长，SE 与 AE 的计算方法见公式(3)与公式(4)。需要说明的是本研究采用 24 小时区间与单位小时内 15 分钟时间间隔作为时间序列串分割的标准，例如 00: 00 – 00: 15 分记为时刻 1，00: 15 – 00: 30 分记为时刻 2，以此类推可以得到 96 个时刻间隔。

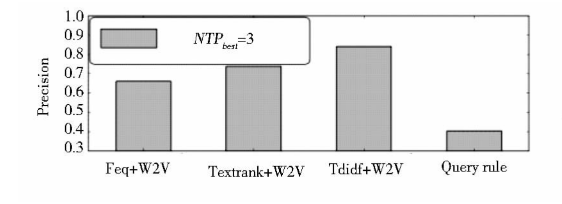
$$SE = -\sum_{j=1}^{Z} P(x_j) \cdot \log(P(x_j)) \quad 式(3)$$

$$AE = \left(\frac{1}{n} \cdot \sum_i \varphi_i\right)^{-1} \cdot \ln(n) \quad 式(4)$$

4　实证结果

4.1　MaxMatching 算法评估

MaxMatching 算法目标是将查询项转换为给定的主题，其算法质量会严重影响后续聚类建模的准确性，并且此方法是基于查询扩展策略所获取的元数据文本进行计算的，因而参数设置会对 MaxMatching 产生直接影响。因此本研究将考虑两个重要的参数：①百度返回的记录数目(NTP)，这里选取的范围为[1, 10]；②元数据的关键词抽取算法(SKE)，本研究纳入考量的是三种常见的文本特征抽取方式，即词频(frequency)、TD-IDF 和 TextRank[43]。另外本研究随机选择 2 000个查询项(query)，并分派给 7 位标引员进行人工标引(人工标引的最大概率类别即为查询项所属主题类别)，同时为了展现 MaxMatching 算法的优势，本研究将一种基于规则匹配的算法[44]作为基准 Baseline 以方便对比。图 4 说明基于"TDIDF + W2V"的 MaxMatching 算法是最优的，其准确率可以达到84.76%，此时最优参数 NTPbest = 3。由此可见，相比于传统的规则匹配方法"query rule"，结合 Word2Vec 深度学习的查询项扩展算法 MaxMatching 在识别用户搜索意图和主题偏好的任务上更高效和准确。



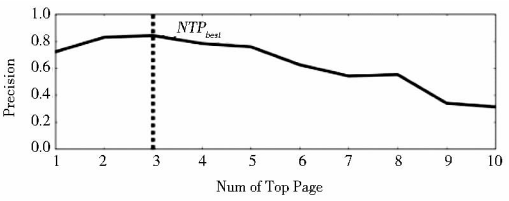
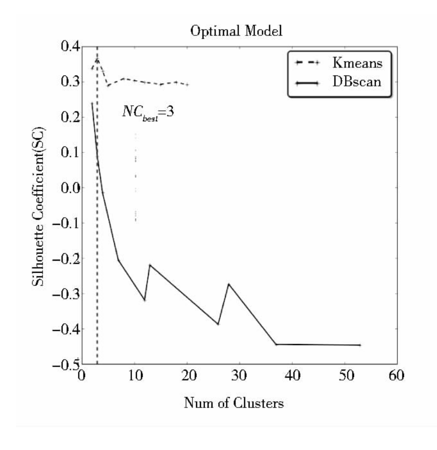


图 4　MaxMatching 算法质量评估

4.2　最优聚类模型

本研究采用两种经典的聚类模型(K-means & DB-scan)进行用户的特征聚类，并根据轮廓系数(silhouette coefficient, SC)[45]进行聚类模型的质量评价，如图 5 所

示。我们对 DBscan 的参数(扫描半径 Eps 以及最小邻居数目 Msn)以及 K-means 的用户聚类数目进行讨论与测试,发现 K-means 算法总体上优于 DBscan,并且当聚类数目 =3 时,K-means 的 SC 值最高,此时聚类效果最佳。



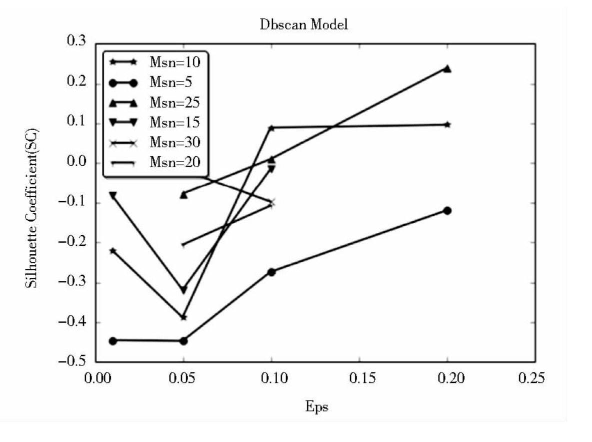


图 5 最优聚类算法和参数评估

4.3 聚类结果分析

通过聚类模型我们可以得到三类不同用户群体(cluster0,cluster1,cluster2)。从样本总体来说,绝大部分大学生用户使用搜索引擎频次不高(73.21% 的 SC_{ratio} 低于均值 0.15,见图 6),近一半(50.79%)的用户基本没使用过搜索引擎。从聚类质心和均值统计来看(见图 6),cluster0 具备较高的特征熵值(SE 和 AE),且 SC_{ratio} 值取值范围几乎均匀涵盖[0,1]区间;cluster1 的 SE 值最低而 AE 值最高,且 SC_{ratio} 处于较低水平范围(95% 的用户 SC_{ratio} 在[0, 0.3]范围内);cluster2 的 AE 值最低,SE 值处于平均水平(SE_{mean} = 0.73),且 SC_{ratio} 处于类似较低水平范围(95% 的用户 SC_{ratio} 在[0, 0.2]范围内)。为了进一步确定上述结果在统计意义上是否显著有效,我们采用 Mann-Whitney U 的非参数秩和检验(由于数据方差不齐),表 2 表明尽管 cluster1 和 cluster0 在 AE 的质心均值上基本相等,但不同两个群体的用户熵特征(SE 和 AE)之间整体上是存在显著差异的。

表 2 Mann-Whitney U 检测

Statistical Indicator	1Mean ± Std	Mann-Whitney U	P-value
Pair | Variable		SE	
(cluster1, cluster0)	0.330 ±0.253	8 068.5	0
(cluster0, cluster2)	1.198 ±0.279	229 625	0
(cluster2, cluster1)	0.721 ±0.466	641 684	0
Pair | Variable		AE	
(cluster1, cluster0)	2.485 ±0.294	1 380 296.5	0
(cluster0, cluster2)	2.339 ±0.323	927 116.5	0
(cluster2, cluster1)	1.371 ±0.401	790 129.5	0

注:"1Mean ± std"为第一个划线的聚类统计指标

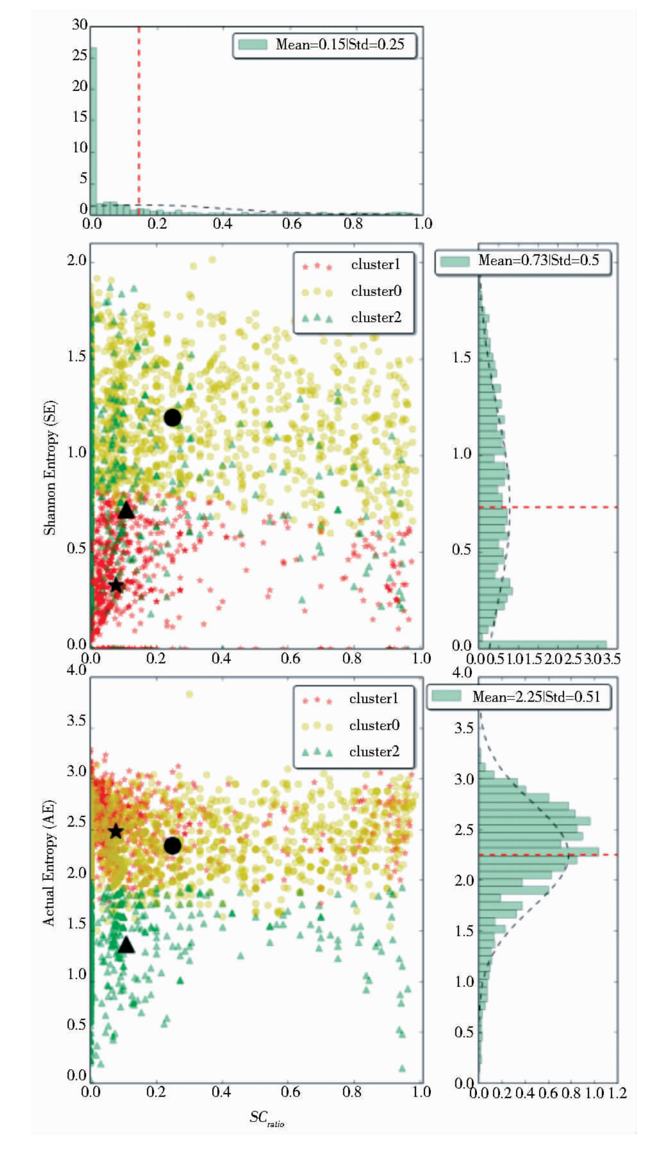


图 6 聚类特征分析

为了更清楚地展示和解释特征熵在用户访问行为的时序性(temporality)和主题性(topicality)上的特点,

我们将这三类用户投影到时钟形式和主题分布进行可视化分析，如图 7 所示。从主题性指标 SE 和时序性指标 AE 的分布来看，三类大学生用户群体存在明显的聚类区分轮廓，这也说明熵特征的区分效果较好。其次，具有最小 AE 值的 cluster2 用户上网最有规律，他们会选择 13：15－13：30，17：15－18：00，19：15－19：30 和 21：45－22：00 的时间段进行上网（绿色部分），然而其他两类用户的在线活跃时间显著长于 cluster2，基本覆盖了 1/3 的整天时间（12：00－2：30 和 16：00－22：30），因此这两类用户的网络访问呈现无序性，即难以预测较为精确的网络所使用时刻，但是值得注意的是 cluster1 群体的平均访问强度（时段均访问次数为 65.3）是明显高于 cluster2（时段均访问次数为38.3）与 cluster0（时段均访问次数为 28.2）。在主题偏好分布方面，cluster1 具有最小 SE 值表现出明显的主题专一性，这类用户对视频直播类（“Live video”）使用和偏好显著高于其他类别（红色部分），相比之下虽然其他两类用户的视频直播类使用量更高，但与其他主题类别的差异性上，这两类用户并不如 cluster1 显著，特别是 cluster0 似乎并没有对某一类主题存在明显的偏好倾向。

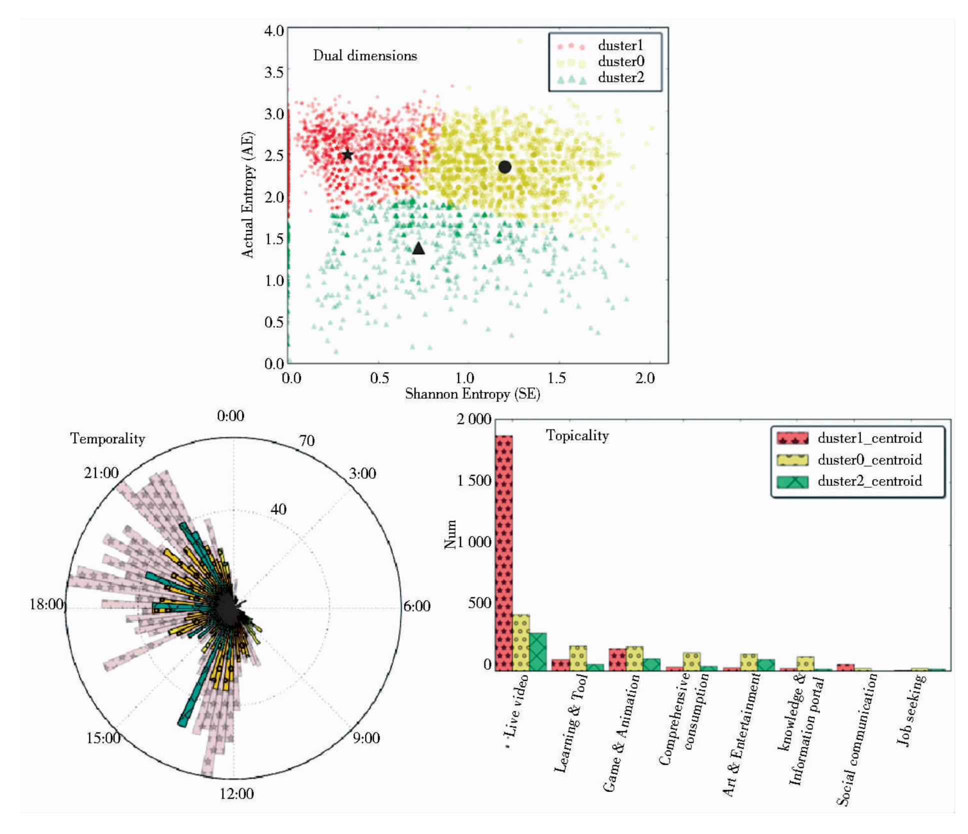


图 7　特征熵的可视化显示

基于上述特征三元组的分析结果，我们归纳出日常网络生活中三类大学生用户群体细分模式，即综合使用型用户（conjoint-utilizing users，CU）、单一使用无序型用户（single-utilizing users in disorder，SUD）和单一使用有序型用户（single-utilizing users in orderness，SUO），分别对应 cluster0，cluster1，cluster2，其具体的群体特征如表 3 所示：

表 3　大学生群体细分特征

大学生用户群	网络使用能力（behavior）	访问时序性（temporality）	主题倾向性（topicality）
CU	综合使用 URL 链接点击与搜索进行信息获取，网络使用能力较强	访问呈现较强的无序性，在线活跃时间持续性长，活跃强度较长	呈现多样性主题偏好，无明显主题倾向性
SUD	以点击行为为主，较少使用搜索工具，网络使用能力较弱	访问呈现较强的无序性，在线活跃时间持续性长，活跃强度较弱	对视频类内容有强烈的单一性倾向
SUO	以点击行为为主，较少使用搜索工具，网络使用能力较弱	访问非常有序和规律，在线活跃时间持续性非常短，活跃强度较弱	呈现多样性主题偏好，对视频类内容有一定较弱的主题倾向性

5 讨论与总结

本研究从行为维度(网络能力)、时间维度(时序性)以及主题维度(主题专一性)三个层面使用中国大学生用户大规模上网行为日志数据构建出一套用户在线行为模式识别的新方法。该方法的核心包括基于深度学习和查询扩展策略的“MaxMatching”匹配算法和不同维度的特征熵测度算法。从算法评测结果来看,相比于传统的规则匹配方法,该算法在识别查询项的用户意图方面表现优异,这对于有效拓展和丰富信息检索领域用户意图理解任务具有一定理论与实践价值,这是研究方法上的贡献。

多维度的特征熵引入可以从全新的角度理解和揭示大学生上网行为模式。实证结果表明:①当大学生用户日常使用综合型导航网站时,他们较少地使用网站内部的搜索工具和组件(如搜索框),而更愿意使用导航链接功能(表现为点击一些热门网站链接),这体现出大学生用户的日常网络生活并不紧密依赖于搜索工具。造成这一“奇怪”现象的原因可能在于大学生用户使用导航网站的意图一般比较简单且目标明确,例如用户希望查询机票和旅游信息时,第一时间会考虑携程和去哪儿,而当他们购买衣装服饰时,很容易就会想到淘宝和京东,他们只需要点击导航网站的链接就可以快速地访问受欢迎的第三方平台,并寻找自己需要的信息资源以满足自己的信息需求,逻辑上一般不会使用更复杂的搜索策略,这符合“省力原则”的解释[46]。②从研究结果来看,本研究基于大规模数据集实验将中国在线大学生用户群体细分为三类群体,包括综合使用型用户、单一使用无序型用户和单一使用有序型用户。尽管视频类网站是大学生用户主要关注的主题类别,三类群体仍具有明显的特征差异——综合使用型用户会充分使用导航链接和搜索查询工具进行信息访问与内容获取,体现出较强的网络使用能力,同时在访问时序性上具有较长的活跃期和较高的活跃强度,但对主题内容层面没有显著的专一性;单一使用无序型用户则以点击热门类网站行为为主,具有较长在线网络活跃时间和较弱的活跃强度,且对视频类信息具有专一性偏好;单一使用有序型用户也是以点击链接为主,但在网络使用时间上非常规律,其活跃时间的长度与强度都比较低,同时他们对主题内容也无明显偏好。总之,这些结论将有助于帮助我们更好地理解大学生这类特殊群体的网络使用行为的模式与特征,从而为针对该特定群体的用户行为所展开的相关研究提供借鉴和参考。

对于广大面向大学生用户的服务商(特别是大部分中小型企业)而言,用户市场细分与用户行为模式挖掘能够有效地帮助企业了解用户群体需求,以支持更为个性化的信息推荐服务,乃至拓展潜在的用户群体和新的服务模式,实现企业数据增值。本研究正是引入了一种基于企业访问日志进行用户市场细分的方法,该方法在数据层面和模型应用层面都较为容易。另外针对这三类用户群体,企业可以制定出符合不同群体的个性化信息推送策略。比如通过此方法可以对识别出的单一使用无序型用户进行长时间的单一类型信息内容推送,信息内容只涵盖“视频直播类”资源即可;但对于单一使用有序型用户,企业应该采用定时混合推荐的策略,即在固定的时钟内(如本研究的4个短时段)进行泛化主题的信息内容推送,内容范围可以涵盖学习工具类、游戏动漫类、文艺娱乐类和视频直播类资源。一方面这种策略有利于较为精确地把握用户群体的定向需求,并作为更精确个性化服务的中间处理环节;另一方面,这种安排可以实现对不同群体的定时定向自动化推送服务,一定程度上提高了企业计算资源利用率,降低服务器不必要的开销和人力维护成本。

然而本研究仍然存在一些不足之处:①实验数据是以构建的虚拟导航平台为基础的,并没有记录完整日常用户网络使用的全部情况,例如用户可能不使用该导航网站而直接使用搜索引擎进行信息查询和使用等,而这类日志数据我们是无法获取的。因此,尽管我们采用绑定导航在网关入口和降低资费政策等方式尽可能地增加用户对平台使用效率,以更好地获得更完整的日志数据,所得到的结果是否一定完全无偏地反映用户搜索工具较低的使用效率等结论仍有待商榷和进一步确认,特别是有关产生该现象的原因如“省力原则”等社会心理因素也需要进一步通过问卷调研和深度访谈予以分析和判断。②虽然相关大学生访问数据集比较少,但本研究使用的数据集以及数据维度仍然需要进一步拓展,以更有力地论证本研究提出的这种方法的有效性和泛化性。我们将在未来工作中继续优化导航网站设计,增加宣传和投放范围,以吸引更多的大学生用户流量。③“MaxMatching”算法依赖于种子词集合的人工标注质量且该算法属于硬性聚类模型,其模型准确率仍需进一步提高。下一步我们将考虑软学习方式,通过加入约束条件和实体识别算法对查询项进行更为精确的识别,同时使用不同的人工标注水平和数量的种子集进行多次重复试验,以实现更好的

模型结果。最后,本研究的实验对象主要是采用 PC 机进行相关测试实验,没有考虑移动端的使用情况,未来的工作将会考虑不同设备途径(如手机端、平板电脑)并结合相应的人口统计特征对上述用户群进行更精细化的特征分析和统计描述,以更全面和深入地挖掘和展示大学生用户在线行为的模式特点和规律。

参考文献:

[1] 中国互联网络信息中心. 第41次《中国互联网络发展状况统计报告》[EB/OL]. [2018－03－05]. http://www. cnnic. net. cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/P020180305409870339136. pdf.

[2] HASSAN M T, KARIM A. Impact of behavior clustering on Web surfer behavior prediction. [J]. Journal of information science & engineering, 2011, 27(6):1855－1870.

[3] JAMALI H R, NICHOLAS D, HUNTINGTON P. The use and users of scholarly e-journals: a review of log analysis studies[J]. Aslib proceedings, 2005, 57(57):554－571.

[4] KINNALLY W, LACAYO A, MCCLUNG S, et al. Getting up on the download: college students' motivations for acquiring music via the Web[J]. New media & society, 2008, 10(6):893－913.

[5] FORTSON B, SCOTTI J, CHEN Y C, et al. Internet use, abuse, and dependence among students at a southeastern regional university[J]. Journal of American college health, 2007, 56(2):137－144.

[6] WANG Y, NIIYA M, MARK G, et al. Coming of age (digitally): an ecological view of social media use among college students [C]//Proceedings of the 18th ACM conference on computer supported cooperative work & social computing. New York: ACM, 2015:571－582.

[7] TENOPIR C. Use and users of electronic library resources: an overview and analysis of recent research studies[M]. Washington, DC: Council on library & information resources, 2003:72.

[8] MADDEN M, RAINIE L. Music and video downloading moves beyond P2P[M]. Washington, DC: Pew Internet and American life project, 2005.

[9] ZHANG P, LIU C. Personal information management practices of Chinese college students on their smartphones[C]// The third international symposium of Chinese CHI. New York: ACM, 2015:47－51.

[10] WU D, Liang S. Research on the follow-up actions of college students' mobile search[C]//Proceedings of the 16th ACM/IEEE-CS on joint conference on digital libraries. New York: ACM, 2016:59－62.

[11] BRODER A Z. A taxonomy of Web search[C]//Proceeding of ACM SIGIR forum. New York: ACM, 2002, 36(2):3－10.

[12] ALONSO O, STONE M. Building a query log via crowdsourcing [C]//Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval. New York: ACM, 2014:939－942.

[13] GONZALEZ-CARO C, BAEZA-YATES R. A multi-faceted approach to query intent classification[C]// Proceedings of the 18th international conference on string processing and information retrieval. Berlin: Springer-Verlag, 2011:368－379.

[14] BAEZA-YATES R, CALDERON-BENAVIDES L, GONZALEZ-CARO C. The intention behind Web queries[C]// Proceedings of the 13th international conference on string processing and information retrieval. Berlin: Springer-Verlag,2006:98－109.

[15] KHUDABUKHSH A R, BENNETT P N, White R W. Building effective query classifiers: a case study in self-harm intent detection [C]//Proceedings of the 24th ACM international on conference on information and knowledge management. New York: ACM, 2017: 1735－1738.

[16] MANSOURI B, ZAHEDI M S, CAMPOS R, et al. Online job search: study of users' search behavior using search engine query logs[C]//Proceedings of the 41th international ACM SIGIR conference on research & development in information retrieval. New York: ACM, 2018:1185－1188.

[17] KANG I H, KIM G C. Query type classification for Web document retrieval[C]//Proceeding of the 26th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2003:64－71.

[18] SUN J, XU J, ZHENG K, et al. Interactive spatial keyword querying with semantics[C]//Proceedings of the 2017 ACM on conference on information and knowledge management. New York: ACM, 2017:1727－1736.

[19] GUO Q, AGICHTEIN E. Exploring mouse movements for inferring query intent [C]//Proceeding of the 31th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2008:707－708.

[20] PUJERI R V, KARTHIK G M. Constraint based frequent pattern mining for generalized query templates from Web log[J]. International journal of engineering science & technology, 2011, 2(11): 17－33.

[21] CAO H, JIANG D, PEI J, et al. Context-aware query suggestion by mining click-through and session data[C]// Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2008:875－883.

[22] TEEVAN J, DUMAIS S T, LIEBLING D J. To personalize or not to personalize: modeling queries with variation in user intent[C]// Proceeding of the 31th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2008:163－170.

[23] CHUANG S L, CHIEN L F. Towards automatic generation of query taxonomy: a hierarchical query clustering approach[C]// Proceedings of the 2002 IEEE international conference on data mining. Washington, DC: IEEE Computer Society, 2002:75－82.

[24] PARK J Y, O-HARE N, SCHIFANELLA R, et al. A large-scale study of user image search behavior on the web[C]//Proceedings of the 33rd annual ACM conference on human factors in computing systems. New York: ACM, 2015:985－994.

[25] LE D T, BERNARDI R. Query classification using topic models and support vector machine [C]// Proceedings of ACL 2012 student research workshop. Stroudsburg: Association for Computational Linguistics, 2013:19 – 24.
[26] ZHAI H, GUO J, WU Q, et al. Query classification based on regularized correlated topic model [C]//Proceedings of the 2009 IEEE/WIC/ACM international joint conference on Web intelligence and intelligent agent technology. Washington, DC:IEEE Computer Society, 2009:552 – 555.
[27] ZHANG C W, FAN W, DU N, et al.. Mining user intentions from medical queries: a neural network based heterogeneous jointly modeling approach[C]//Proceedings of the 25th international conference on World Wide Web. The Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016:1373 – 1384.
[28] HASHEMI S H, WILLIAMS K, KHOLY A E, et al. Measuring user satisfaction on smart speaker intelligent assistants using intent sensitive query embeddings[C]//Proceedings of the 2018 ACM on conference on information and knowledge management. New York: ACM, 2018:1183 – 1192.
[29] DOU S, SUN J T, YANG Q, et al. Building bridges for web query classification [C]// Proceeding of the 29th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2006:131 – 138.
[30] KONISHI T, OHWA T, FUJITA S, et al. Extracting search query patterns via the pairwise coupled topic model[C]//Proceedings of the 9th ACM international conference on Web search and data mining. New York: ACM, 2016:655 – 664.
[31] 郭程, 白宇, 郑剑夕, 等. 一种无指导的子主题挖掘方法[J]. 中文信息学报, 2016(1): 50 – 55.
[32] WU B, XIONG C Y, SUN M S, et al. Query suggestion with feedback memory network [C]//Proceedings of the 27th international conference on World Wide Web. The Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2018:1563 – 1571.
[33] WANG Z, WANG F, WANG H, et al. Unsupervised head-modifier detection in search queries[J]. ACM transactions on lnowledge discovery from data, 2016, 11(2):1 – 28.
[34] DUAN H, ZHAI C X. Mining coordinated intent representation for entity search and recommendation [C]//Proceedings of the 24th ACM international on conference on information and knowledge management. New York: ACM, 2015:333 – 342.
[35] 冯晓华, 陆伟, 张晓娟. 检索结果多样化研究综述[J]. 情报学报, 2015, 34(7):776 – 784.
[36] LIU P Q, AZIMI J, ZHANG R f, et al. Contextual query intent extraction for paid search selection[C]//Proceedings of the 24th international conference companion on World Wide Web. New York: ACM, 2015:71 – 72.
[37] SEOCK Y K, CHEN Y. Website evaluation criteria among US college student consumers with different shopping orientations and Internet channel usage[J]. International journal of consumer studies, 2007, 31(3):204 – 212.
[38] MALIK A, MAHMOOD K. Web search behavior of university students: a case study at university of the Punjab[J]. Webology, 2009, 6(2):1 – 13.
[39] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of the 26th international conference on neural information processing systems. New York: Curran Associates Inc. ,2013: 3111 – 3119.
[40] TABATABAI D, SHORE B M. How experts and novices search the Web[J]. Library & information science research, 2005, 27(2): 222 – 248.
[41] SAVOLAINEN R. Network competence and information seeking on the Internet: from definitions towards a social cognitive model[J]. Journal of documentation, 2002, 58(2):211 – 226.
[42] SONG C, BARABASI A L. Limits of predictability in human mobility[J]. Science, 2010, 327(5968):1018 – 1021.
[43] MIHALCEA R. TextRank: bringing order into texts[C]//Proceeding of 2004 conference on empirical methods in natural language processing. Barcelona: ACL, 2004: 404 – 411.
[44] KOUTRIKA G, IOANNIDIS Y. Rule-based query personalization in digital libraries[J]. International journal on digital libraries, 2004, 4(1):60 – 63.
[45] ROUSSEEUW P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of computational & applied mathematics, 1999, 20(20):53 – 65.
[46] 王晓娜. 最小省力原则与情报检索系统的可接近性[J]. 情报科学, 2000, 18(2):135 – 136.

作者贡献说明:
严承希:论文撰写、实验编码和数据分析;
王军:思路提出与论文修改;
王珂:实验评估与论文修改。

Chinese College Students' Internet Use: A New Method of Behavior Pattern Recognition with Massive Log Analysis

Yan Chengxi Wang Jun Wang Ke
Department of Information Management, Peking University, Beijing 100871

Abstract: [**Purpose/significance**] It is of great significance to analyze and understand users' daily Web behavior

patterns, which not only makes progress in the domain of user behavior analyse and information retrieval theoretically, but also has potential social values and practical significance in promoting personalized service and information recommendation for the undergraduate-oriented enterprises. [**Method/process**] In this paper, a new method for college students' behavior Web pattern recognition based on large-scale log analysis was proposed. It included a semi-supervised learning algorithm "MaxMatching" based on deep learning and text analysis, and a hybrid model combined with two characteristic entropy (Shannon Entropy and Real Entropy). [**Result/conclusion**] The empirical results showed that this method has the excellent performance in the algorithm and the result interpretation. Also, it can generalize and present all-round Chinese college students' Web behavior pattern in three aspects of network ability, temporality and topicality. The method and conclusion can effectively expand the methods about semantic understanding of queries in information retrieval, and provide some reference and feasible suggestions to undergraduate-oriented enterprises on personalized recommendation service.

Keywords: Chinese students online behavior pattern recognition massive log analysis

情报学与情报工作发展论坛(2019)征稿通知(第一轮)

情报学与情报工作发展论坛自成立以来,已成功举办两届,有效推动了情报学与情报工作的科学发展,并取得了良好反响与广泛肯定。大数据与人工智能正在重塑情报学与情报工作的内核与应用场景,为延续《南京共识》精神,把握转型与变革机遇,汇集并凸显情报领域的最新进展,推动我国情报学人与情报工作者的交流,创新情报学与情报工作的理论与实践,搭建年度性的全国情报学学术会议平台,形成学术传统,"新时代 新使命 新作为——情报学与情报工作发展论坛(2019)"将于2019年11月8日-10日在武汉华中师范大学举办。本次论坛将秉承情报学与情报工作发展论坛优良传统,邀请地方、军队、公安等高校和军队、地方情报所的专家学者、师生代表、从业人员共同参会,围绕新时代情报学与情报工作创新与发展展开深入的交流和碰撞,通过不同领域学者专家的探讨与互动,推动情报学与情报工作的纵深发展。热忱欢迎情报学与情报工作领域的师生、学者、专家、从业人员踊跃投稿并参会!

一、主办单位

中国科学技术情报学会
中国社会科学情报学会
中国国防科学技术信息学会
华中师范大学信息管理学院

二、会议日期

2019年11月8日-10日

三、会议地点

武汉·华中师范大学

四、征稿主题:新时代情报学与情报工作创新与发展

本届论坛征稿主题包含但不限于以下主题,供投稿作者选题参考。

(1)情报学理论发展与创新。
(2)情报学方法创新与应用。
(3)情报技术创新与实践。
(4)信息行为与情报服务。
(5)安全情报。
(6)情报学学科建设。
(7)情报工作与情报事业发展。

五、征稿要求

(一)征稿对象

论坛面向情报学与情报工作领域的师生、学者、专家、从业人员征稿。

(二)重要日期

征文截稿日期:2019年8月31日
审稿结果通知:2019年9月30日
稿件请发送至论坛专用邮箱:qbxqbgz2019@163.com

(三)稿件要求

投稿论文须是未公开发表的原创性研究成果,篇幅字数控制在8 000字左右。投稿论文格式请参照《图书情报工作》期刊的"投稿须知及格式规范"。

(四)录用、评奖与发表

论坛主办方将邀请专家对投稿论文进行严格评审,一经录用酌付稿酬,并为受邀作论文交流的作者提供与会期间的食宿(每篇录用论文限资助一位);根据征稿数量和质量从中评选出优秀论文一、二、三等奖,届时颁发荣誉证书与奖励;优秀论文将推荐给《图书情报工作》、《图书情报知识》、《情报学报》、《情报科学》、《情报理论与实践》、《信息资源管理学报》、《情报工程》、《情报杂志》、《现代情报》、《知识管理论坛》、《农业图书情报》(排名不分先后)等期刊发表。

六、联系方式

华中师范大学信息管理学院 李玉海
邮箱:yhli@mail.ccnu.edu.cn
电话:027-67868865
华中师范大学信息管理学院 易明
邮箱:yiming0415@mail.ccnu.edu.cn
电话:13387599231

特此通知。

华中师范大学信息管理学院
情报学与情报工作发展论坛(2019)组委会
二〇一九年四月二日